兰州大学

本科毕业论文

论文题目（中文） 基于 Deep Koopman 算子网络的非线性系统强化学习研究

论文题目（英文） Deep Koopman Network Based Reinforcement Learning of Nonlinear System

学生姓名 许F欢

指导教师 赵东东

学　　院 信息科学与工程学院

专　　业 电子信息科学与工程

年　　级 2020 级

兰州大学教务处

诚信责任书

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人、集体已经发表或未发表的论文。

本声明的法律责任由本人承担。

论文作者签名：签名　　日　　期：签名

关于毕业论文（设计）使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用毕业论文（设计）的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业论文（设计）。本人离校后发表、使用毕业论文（设计）或与该毕业论文（设计）直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本毕业论文（设计）研究内容：

☑可以公开

□不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名：签名　　导师签名：签名

日　　期：签名　　日　　期：签名

基于 Deep Koopman 算子网络的非线性系统强化学习研究

中文摘要

我的摘要

关键词： Koopman 算子理论，深度神经网络，强化学习

Deep Koopman Network Based Reinforcement Learning of Nonlinear System

Abstract

My Abstract

Keywords: Koopman Operator Theory, Deep Neural Network, Reinforcement Learning

目　录

第一章 绪 论

这是我的绪论 [1]

第二章 背景知识

在本章中，首先讨论一下有关的背景理论与算法。介绍一下 Koopman 算子理论（Koopman Operator Theory），并讨论 Koopman 算子对于重塑强化学习（Reinforcement Learning）中使用的马尔可夫决策过程（Markov Decision Process）的重要作用。同时，对于 Koopman 算子理论与深度神经网络（Deep Neural Network）之间的关联。

2.1 Koopman 算子理论

系统的强非线性是数据驱动建模和控制领域的核心问题之一，包括基于现代强化学习框架所做的工作。Koopman 算子理论 [cite] 为上述问题提出了一种解决方案。在该理论中，非线性系统动力学可以在提升到无限维度希尔伯特空间时变为线性系统，并可以通过一组新的基底对于该线性系统进行观测。升维工作已被证明对于线性化和简化某些具有挑战性的问题具有显著效果，这与机器学习领域中其他的类似努力相一致。

2.1.1 非线性系统描述

我们应该采用不同的形式对不同类型的系统中系统状态 x 进行描述。假设系统为确定性自治系统，我可以采用下面的方式，将一个离散动力系统描述为

$$\dot{x} = F(x) \tag{1}$$

其中，$\dot{x}$ 表示下一个时刻的系统状态；或者，采取不同的方式，将一个连续动力系统描述为

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) \tag{2}$$

而在处理实际问题是，我们通常考虑的是离散动力系统。所以我们就可以将连续系统通过流映射算子（Flow Map Operator）归纳为一个离散系统，系统演化如下

$$x(t+\tau) = F_\tau(x(t)) = x(t) + \int_t^{t+\tau} f(x(s))\mathrm{d}s \tag{3}$$

2.1.2 Koopman 算子与系统演化

Koopman 算子提供了解决非线性系统控制问题的一个新的着眼点。在形式上，我们考虑一个实值向量测量函数 $g : M \to \mathbb{R}$，且都由无限维希尔伯特空间的元素组成，其中 M 是一个流形。通常，这个流形被认为是 $L^\infty(X)$, $X \subset \mathbb{R}^d$。一般情况下，函数 g 被称为可观测函数。Koopman 算子理论中指出，Koopman 算子 $\mathcal{K}$ 和 Koopman 生成器 $\mathcal{L}$ 都是作用于上述

观测函数 g 的无限维线性算子，在确定性系统中，有

$$\mathcal{K}g = g \circ F \tag{4a}$$

$$\mathcal{L}g = f \cdot \nabla g \tag{4b}$$

Koopman 生成器 $\mathcal{L}$ 和 Koopman 算子有如下的关系

$$\mathcal{L}g = \lim_{t\to 0} \frac{\mathcal{K}g - g}{t} = \lim_{t\to 0} \frac{g \circ F - g}{t} \tag{5}$$

Koopman 算子理论可以更广泛地应用于任何马尔可夫过程，但本文以随机性连续时间系统为例，此时，Koopman 算子的定义如下:

$$\mathcal{K}g = \mathbb{E}(g(X)|X_0 = \cdot)$$

$$\mathcal{L}g = \lim_{t\to 0} \frac{\mathcal{K}g - g}{t}$$

Koopman 算子将测量函数 g 沿着路径 x 向前演化如下：

$$\mathcal{K}_\tau g(x_t) := g(F_\tau(x_t)) = g(x_{t+\tau}) \tag{6}$$

其中 F 代表着系统的演化规律，或者更一般的, 在随机自治系统中，F 被如下定义为条件预测算子：

$$\mathcal{K}g(x_t) = \mathbb{E}\left[g(X_{t+\tau})|X_t = x_t\right] \tag{7}$$

在上述离散系统算子中，普遍使用 $\mathcal{K} := \mathcal{K}_1$，在本文中也照此用法。

2.1.3 Koopman 算子本征函数

上文提出，可观测函数 g 的观测，都存在于无限维的希尔伯特空间中（被称作观测空间），被如 Koopman 算子 $\mathcal{K}$ 等无限维的算子推动沿着给定非线性动力学系统演化。因此不难发现，我们可以应用 Koopman 算子理论研究，将对于非线性系统的研究，通过观测空间状态线性演化实现。

同时，由于难以捕捉无限维希尔伯特空间中所有可观测函数的演化，所以应该试图识别随着非线性系统动力学而线性演化的关键观测函数，Koopman 算子的本征函数就可以作为一组特殊的观测函数 [cite]：

$$\mathcal{K}\Phi(x_k) = \lambda\Phi(x_k) = \Phi(x_{k+1}) \tag{8}$$

此时，本征函数就会成为观测空间的一组基底，由此，数学上，观测函数应当表示为这组基底的线性组合，如下：

$$g(x) = \sum_{i=1}^{n} a_i \Phi_i(x) \tag{9}$$

当前，从数据中挖掘信息并获得 Koopman 本征函数是现代动力学系统研究的主流方法，被称为数据驱动（Data-Driven）的 Koopman 算子。由此，我们可以通过数据驱动的方式得到 Koopman 算子的本征函数，以得到非线性系统在高维空间中的全局线性表示。

此外，Koopman 算子已经被广泛运用于受控系统。在受控确定性离散时间系统中，我们有：

$$x' = F(x,u) \tag{10}$$

在受控连续时间系统中：

$$\frac{\mathrm{d}}{\mathrm{d}t} = f(x(t),u(t)) \tag{11}$$

2.2 强化学习

强化学习（Reinforcement Learning）是机器学习（Machine Learning）和控制理论（Control Theory）的交叉领域，在强化学习中，智能体学习如何与复杂环境进行交互，并以获得更高的奖励为目标。最近，深度强化学习（Deep Reinforcement Learning）被证明能够在若干项具有挑战性的任务中，实现人类水平或超人类的表现，包括玩电子游戏 [cite] 和策略游戏 [cite]。深度强化学习也越来越多地用于科学和工程应用，包括药物发现 [15]、机器人操作 [16]，自动驾驶 [17] 和无人机赛车 [18]、流体流量控制 [19-25] 和融合控制 [26]。

2.2.1 马尔可夫决策过程与贝尔曼方程

马尔可夫决策过程是具有马尔可夫性质的随机过程。

随机过程是指研究对象是随时间演变的随机现象，在随机过程中，随机现象在某一时刻 t 的取值是一个向量随机变量，可以用 S_t 表示，所有可能的状态组成状态空间 $\mathcal{S}$。我们将已知所有历史状态 $(S_1,\ldots,S_t)$ 时，某一时刻 t 的状态 S_t 发生的概率用 $P(S_{t+1}|S_1,\ldots,S_t)$ 表示。而马尔可夫性质则表示，已知当前时刻状态 S_t 时，下一时刻状态 S_{t+1} 仅与 S_t 有关，用 $P(S_{t+1}|S_t)$ 表示。而从前一状态经过随机进入下一状态的过程被称为状态转移。

在下文中，我们假设存在一个无限时域的马尔可夫决策过程。我们假设代理跟随随机策略 $\pi(u|x)$，表示在已知状态 x 的情况下，采取某个特定动作 u 的可能性。由此，在离散时间系统中，状态价值函数定义为：

$$V^{\pi}(x) = \mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r(x_k,u_k)|\pi,x_0=x\right] \tag{12}$$

其中，$\gamma\in[0,1]$ 表示折扣率，$r(x_k,u_k)$ 表示智能体得到的奖励。

在这里我们强调强化学习与控制领域的结合，所以本文中考虑使用线性二次最优控制问题。线性二次最优控制是十分经典的控制领域问题，其中线性代表研究的系统动态可以用一组线性微分方程表示，而其成本为二次泛函。形式上，我们考虑有限时间长度，离散

时间的 LQR，假设离散时间线性系统：

$$x_{k+1} = Ax_k + Bu_k$$

其性能指标为：

$$c(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k \tag{13}$$

将系统线性二次性能指标的相反数作为智能体探索时的奖励:

$$r(x_k, u_k) = -c(x_k, u_k) \tag{14}$$

并改写贝尔曼方程可以得到：

$$V^\pi(x) = \mathbb{E}\left[\sum_{k=0}^{\infty} -\gamma^k c(x_k, u_k)|\pi, x_0 = x\right] \tag{15}$$

2.2.2 最大熵强化学习

强化学习智能体在马尔可夫决策过程学习策略的过程中，主要需要面对的问题有两个，

参考文献

[1] Tenne R, Margulis L, Genut M e, et al. Polyhedral and cylindrical structures of tungsten disulphide[J]. Nature, 1992, 360(6403):444–446.

附　　录

这里是我的附录这里是我的附录
这里是我的附录

致　　谢

这里是致谢页

(我是谁？兰朵儿开发者：余航，致谢我，查重时一定会重复的，哈哈，开个玩笑，本科生论文不在查重范围，而且“毕业论文(设计)检测内容主要为毕业论文(设计)的主体部分”)。

毕业论文（设计）成绩表

导师评语

好好好

建议成绩____签名____ **指导教师（签字）**____签名____

答辩委员会意见

优秀

答辩委员会负责人（签字）_____

成绩____**100**____ **学院（盖章）**_____

年 月 日